DOCUMENT RESUME

ED 212 324 JC 820 015

AUTHOR Yess, James P.
TITLE Should College Faculty Be Rated by Students Using Instruments Developed at the Bargaining Table?
PUB DATE [81]
NOTE 34p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Behavior Rating Scales; Case Studies; *Collective Bargaining; College Faculty; Community Colleges; Observation; *Reliability; *Student Evaluation of Teacher Performance; Student Teacher Relationship; Two Year Colleges

ABSTRACT

Instruments used to rate college teachers should undergo the scientific scrutiny expected of any psychological instrument. The construction of rating instruments through collective bargaining without scientific pilot testing thwarts the desired goal of assisting faculty and administrators in learning about teaching performance. These contentions are supported by a brief discussion of the divergent goals of psychological rating and collective bargaining. The argument is backed by conclusions drawn from a case study of a rating instrument developed in the late 1970s during collective bargaining sessions between faculty union representatives and governing body representatives concerning working conditions at 15 community colleges. An agreement was reached whereby student evaluation, evaluation of course materials, classroom observations, and a composite category of student advisement, college/community service, and professional development, could be used in the process of evaluating faculty for reappointment, tenure, promotion, multiple-year contracts, and/or termination. Student evaluation was given a weight of 40% and the other three criteria each 20% in this process. The 12-item rating instrument contained serious flaws, notably its ambiguity and lack of scientific testing. The paper concludes with recommendations for developing a sound rating instrument within the context of labor/management negotiations. (Author/HB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED212324

SHOULD COLLEGE FACULTY BE RATED BY STUDENTS USING INSTRUMENTS DEVELOPED AT THE BARGAINING TABLE?

BY

JAMES P. YESS

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

James P. Yess

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☒ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

JC 820 015

SHOULD COLLEGE FACULTY BE RATED BY STUDENTS
USING INSTRUMENTS DEVELOPED AT THE BARGAINING TABLE?

Abstract

The argument presented in this paper advances the view that instruments used to rate college teachers should, prior to implementation, undergo the scrutiny expected of any psychological measurement. Efforts to construct a rating instrument such as this at the bargaining table without regard for pilot testing thwarts the desired goal of assisting faculty and administrators in learning about teaching performance. These contentions will be supported by a brief discussion of the divergent goals of psychological rating and collective bargaining. A case study will also be provided as an example of the problems inherent in ignoring a scientific approach to the development of a teacher rating instrument. Recommendations are presented concerning procedures for developing a sound rating instrument within the context of labor/management negotiations.

THE AUTHOR IS THE ASSISTANT DEAN OF CONTINUING EDUCATION AT MASSASOIT COMMUNITY COLLEGE, BROCKTON, MA.

The presentation in this paper involves brief discussions of the behavioral measurement tool known as rating instruments and the dynamics of faculty collective bargaining. It will be argued that the central goals of collective bargaining and the scientific approach to rating faculty performance are distinctly divergent. A case study will be offered to demonstrate why the title question should be answered in the negative. Faculty should not be rated by their students through the use of an evaluation instrument solely developed at the bargaining table.

RATING INSTRUMENTS

Essentially, the behavioral measurement device known as a rating instrument permits the systemtaic communication of one person's perceptions about himself or another individual to either a third party or to the participants in the rating process. Rating scales generally consist of a list of descriptive behaviors or characteristics (stimulus variables) which the observer judges with respect to how well each item describes the subject. Different response formats have

been devised, such as: multiple-choice, forced-choice, single and double continuum, checklist, ranking, frequency of occurrence, free response, man to man, and percentage of group. Each response format has its merits and disadvantages. Psychologists spend a great deal of effort attempting to improve validity and reliability while attempting to eliminate all sources of error.

These errors which arise due to the instrument itself, the subjects, or the circumstances under which the rating takes place can seriously compromise the information gleaned from the rating procedure. In constructing a rating instrument psychologists take pains to avoid ambiguity when describing the behaviors to be judged. An item such as "Promptly returns homework and tests" appearing on faculty evaluation tool invites various interpretations of "promptly". For some students "promptly" might mean "the very next class" while others may be inclined to think that instructors are prompt when they correct and return student work within two weeks. This ambiguity may appear not only in the stimulus variables but also in the response format. Given response choices on a continuum such as: "almost always...frequently... sometimes...infrequently...almost never", one can see

how much the responses are subject to individual interpretation on the part of the raters.

Raters, themselves, bring with them certain human foibles that must be anticipated when devising an evaluation instrument. Among these idiosyncrasies are: an unwillingness to carefully weigh judgements (perfunctoriness); a tendency to use the same response choice for most scale items (response-set); one or two outstanding qualities (favorable or unfavorable) of the subject influencing the rater's entire judgement on all items of the rating instrument ("halo" effect); the willingness to rate familiar subjects favorably (leniency); and avoidance of rating persons at the extremes of response choices (acquiescence). One can attempt to reduce the effects of these sources of error by a careful analysis of all items in the rating instrument and the judicious choice of response format. As Downie succinctly stated, "In general these scales, like test items, require thought, care, and revision". (4:343) There is, however, one deficiency in the rater that those critics of student evaluations of faculty have repeatedly brought to the fore. A basic shortcoming of all students is that they are un-

trained observers who, according to the critics, fail to use any consistently relevant criteria while observing the classroom performance of their instructors. Without some sort of training for the observers, it is argued that they will lack a uniform standard of reference, will be unable to overcome their biases, and will fail to attend to the pertinent aspects of the behaviors that are being evaluated. Another factor mitigating against accuracy in the instrument or rater is the attributes to be rated and the set of conditiions under which the subject is observed. In constructing the items for rating instruments, it is important to ask the observer to judge the overt behavior of the subject. Items which appear on a faculty evaluation form such as "Appears interested in you as a person" require the observer to broadly infer a covert quality from a myriad of overt behaviors. The danger here is that this inferential leap attempts to span a wide chasm and for that reason is subject to gross inaccuracy. The difficulty of reporting on these covert aspects is compounded by the limits of the interaction available during the observation sessions. These limits

are functions of the frequency of observations, the quality of those observations, and the variety of observational settings. It would appear that the more frequent the observations the more reliable the rating. What should also be considered, however, is how much time the observer actually spent engaged in the process of evaluation and how many different settings was the subject being observed. While in the case of students evaluating their college professors the total time potentially available for observation is substantial, the student for the most part is not occupied with applying criteria for good teaching but rather is occupied with taking notes, and trying to understand lecture material. Effective contact, therefore, between student and instructor for the purposes of rating the instructor's performance is far less adequate than it appears. Students, also, observe their teachers under the restrictions imposed by the classroom setting and the professor's pedagogical preferences. There is, after all, only so much you can learn about teachers who conduct their classes in a traditional lecture mode. Hence, students in this kind of setting have no basis for responding to an

item such as, "Allows for personal growth in the learning process." It is obvious that when the development of a rating instrument is undertaken, extreme care should be exercised to avoid the many pitfalls open to this form of information gathering. Miller (6) emphasizes the dangers involved with taking an unscientific approach to rating faculty. When such ratings are used for administrative purposes, he believes the institutions are duty bound to demonstrate the validity and reliability of these measurement instruments. Miller supports his position by drawing on the results of research to list six reasons why colleges cannot allow students to evaluate faculty willy-nilly. Research appears to provide evidence for Miller's contentions that student ratings of faculty are influenced by: (1) class size, having larger classes may produce lower ratings; (2) differences in subject matter taught; (3) the number of college courses a student has completed prior to rating an instructor, the greater number of courses the tendency to give higher evaluations; (4) the reason for the student's choice of courses, i.e. required course or elective; (5) the demands placed on the

student; (6) the inability of the student to discriminate between effective and ineffective teaching. The development of instruments used by college students to evaluate their professors is deserving of the diligent efforts needed to produce a sound evaluative tool. There must be scientific evidence to assure those who wish to learn something from this exercise that the characteristics assumed to be rated are actually being rated and that the measurement taken records those qualities of performance indicative of the subject's traits rather than momentary digressions from the true features of a person's teaching style. Far from being unreasonable, it is necessary to demand before implimentation that any tool used to evaluate teaching effectivenesss come under the rigorous standards applicable to other forms of psychological testing. Hence, the necessity for conducting pilot studies and using appropriate statistical analyses in order to discover the flaws contained in a proposed instrument. This in turn will lead to fruitful avenues of revision culminating in a more accurate measuring device.

Thorndike and Hagen (9) in addition to Cronbach (3) suggest strategies for refining rating instruments. The following guidelines for constructing and revising a rating instrument include their advice:

(1) establish clear evaluating criteria

(2) avoid single item scales but insure that items of one scale actually measure the same trait

(3) test for ambiguity

(4) use appropriate statistical methods to determine the degree of validity and reliability achieved

(5) determine the ability of the instrument to differentiate between persons, i.e. does the instrument produce a normal distribution

(6) items should measure specific overt behavior

(7) free the instrument from theoretical constructs

(8) adjust the response format so that it is appropriate and efficacious

(9) reconstruct items and the response format so that sources of error due to rater bias are reduced as much as possible

(10) educate and train the rater in the use of the rating tool

(11) use raters who have had an optimal opportunity to observe the subjects

THE GOALS OF COLLECTIVE BARGAINING

The collective bargaining table is not a laboratory bench a researcher approaches with a view toward arriving at truth. The parties to collective bargaining would, indeed, fail to appreciate their activity entirely if they approached their problems from a scientific orientation.

The central aim of collective bargaining is to reach an agreement whose terms are satisfactory to the constituencies represented. These terms deal with such issues as compensation, work-day/work-week, and other conditions of employment. One of these working conditions is the procedure whereby employees are evaluated for retention, promotion, merit increments, or tenure. Thus, since such weighty employment issues hang in the balance of the evaluation procedure, the parties at the bargaining table take great pains in arriving at an evaluation procedure and

setting down the language which describes it. The evaluation section of the collective bargaining contract, however, cannot be divorced from the entire package and, therefore, cannot be criticized apart from other sections which impact or are impacted by the evaluation process.

The sides at the bargaining table want to embody the will of their respective constituencies in the final project of negotiations - the contract. Negotiating team members, therefore, view their mission as one of obtaining the best possible concessions from an adversary seeking opposing concessions. The entire agreement becomes an organic whole growing out of the dynamics of collective bargaining. It is little wonder, then, that after the negotiators have reached an agreement and return to their constituents to interpret its terms prior to a ratification vote, there is a prevailing belief that the process of compromise leaves no one ecstatic but is merely designed to satisfy the majority. Sacrifices have been made on both sides in order to reach agreement. When one side relinquishes a demand concerning a particualr section of the contract, it expects to receive concessions somewhere

else. Consequently, it is difficult to judge the parts of a collective bargaining agreement without understanding the overall interplay within an entire context.

Although the substance of the collective bargaining varies from one type of employment to another, negotiations taking place in higher education, according to Carr and Van Eyck (1), have mainly used the industrial bargaining model. On management's side in higher education the principals represented at the bargaining table are the Board of Trustees and President. In public institutions, while not principals, the subtle influence of the state legislature is ever present. Secondarily, but not without power, are the Deans who must be satisfied to some degree. Negotiating for labor is the faculty negotiating team. This team acts on behalf of the union's governing board which in turn receives its power from the membership comprised of professionals coming form diverse academic disciplines. The forces of the power blocks converge at the bargaining table. Hence, by necessity, negotiations are highly political, emotional, and utilitarian. This is not to say that the negotiation sessions are back-

room free-for-alls where something akin to the law of the jungle prevails. On the contrary, there is a concerted effort on both sides to establish respect and credibility in a civil atmosphere. Elements such as evidence and sound reasoning play a role in the negotiating process but are subordinate to the main goal of reaching an agreement through the give and take characteristic of compromise. Carr and VanEyck (1:158-159) aptly describe the situation:

> "Both sides come to the table armed with 'facts' that each is prepared to interpret and apply to its advantage in trying to reach agreement on specific points. But, since 'the objective of bargaining is agreement...the end may be consummated in spite of the facts'.(8:511) Both sides have emotional and political needs. Each must 'win' some points, even where the facts and analysis may support the other side's contention... For either party to go beyond a certain point in its assurance about its facts, its adherence to its 'principles', and its insistence on its demands, ensures the failure of the process of collective bargaining as a viable means of reaching

agreement on difficult issues, about which, if men insist, they can continue to disagree indefinitely."

One does not condemn those who engage in collective bargaining because it is a quest for agreement rather than truth. The economic, academic, and political problems encountered in negotiations are inordinately complex and to seek their resolution through the scientific method would be inappropriate folly. Equally foolish, however, is to undertake the development of a student rating instrument of faculty at the bargaining table since honestly engaging in this activity necessitates scientific inquiry. Without taking appropriate precautions rating instruments can have disastrous consequences for professors and administrators. Imagine a community of scholars allowing a rating instrument to be used in an evaluation process which helped to determine their employment status unaware of its ability to accurately measure teaching performance. Equally, imagine college presidents and boards of trustees considering the results of these potentially bogus evaluations in order to insure excellence in teaching! Although of more than 400 liberal

arts colleges surveyed by Seldin et al (7) the number of liberal arts colleges using administrative rating forms to evaluate faculty has substantially increased (from 23.9% in 1966 to 54.9% in 1973) and the number of these colleges reporting that they did research on the validity of the rating forms also increased (from 0.6% in 1966 to 8.3% in 1973), the fact remains that fewer than 35 of the 410 colleges responding were moved to make some methodical determination of the instrument's efficacy. It is recognized that these data concern administrative rather than student ratings of faculty and it is not known how many of these rating forms were negotiated at a bargaining table. Nevertheless, the indication remains that collegiate educators may not be very interested in learning about how accurately their ratings measure teaching performance.

The following case study exemplifies the difficulties inherent in adopting a student rating instrument of college faculty without the benefit of pilot studies, validity assessment, and an analysis of reliability. The fact that an instrument such as this was formulated at the bargaining table further weakened its chances of being informative due to the dynamics of labor/

management negotiations. It is hoped that the case study presented here will support the contention that an instrument used to measure a complex set of human behaviors known as teaching cannot adequately be developed through the process of collective bargaining because these distinctly different activities are at cross-purposes.

A CASE STUDY

The rating instrument presented in this section was developed in the late 1970's during collective bargaining sessions that continued for the better part of a year. Items for negotiation centered around working conditions at 15 publicly supported community colleges. Participants in the negotiations were faculty union representatives and representatives of the governing board. Which side initially proposed the issue of a student rating instrument of teaching performance is not known, but the final document was the product of the collective bargaining process. No validity or reliability assessments were performed either through pilot testing prior to system-wide implementation or as an official post-administration evaluation. According to the contract, the results

of these student ratings of tenured and non-tenured full-time faculty could be used in the evaluation process for reappointment, promotion, tenure, multiple-year contracts, and/or termination. The criteria which Division Chairperson, Deans, and Presidents employed for making their respective recommendations consisted of a review of 4 categories of evidence: and evaluation by students; an evaluation of course materials; classroom observation; and an evaluation of student advisement, college/community service, and professional development. The faculty members composite evaluation was determined by giving student evaluations a weight of 40%, course materials 20%, classroom observation 20%, and the remainder 20%.

The instrument used by students to rate their professors consisted of 12 items to which the students applied a frequency of occurrence scale. The scale is comprised of five possible responses designated by numerical indicators. Students make responses to the instrument by writing the apporporiate numeral in a box provided to the left of each of the 12 items. The choices available were: 5 = almost always, 4 = frequently, 3 = sometimes, 2 = infrequently, 1 = almost never.

The 12 items are listed below and the reader will perhaps recall that some of these items were introduced earlier to demonstrate some common flaws of rating instruments.

1. ☐ Presents class material in a way that you understand.
2. ☐ Meets and helps you when it is needed.
3. ☐ Assists you with problems when asked.
4. ☐ Communicates to you the course objectives.
5. ☐ Allows for personal growth in the learning process.
6. ☐ Appears interested in you as a person.
7. ☐ Holds classes as scheduled.
8. ☐ Promptly returns homework and tests.
9. ☐ Indicates areas for improvement when returning assignments.
10. ☐ Present specific course organization, objectives and requirements to students.
11. ☐ Evaluates students according to objectives and requirements of the course.
12. ☐ Uses equitable and just grading and/or evaluation methods.

This form was distributed to students in their classes by faculty members. No instructor was permitted to either distribute or proctor the student evaluation in his/her own class. Instructions to the students appeared on the reverse side of the form. Students were informed that unless they signed the form in the space provided, that their appraisal would not be included in the tabulation of evaluation results. The form was perforated so that when the student returned the completed and signed form to the proctor, the signature portion bearing a five digit number was detached from the evaluation portion bearing the identical number. Those in the data processing center responsible for key-punching the data only saw the evaluation portion of each evaluation. Students were informed that "For the purposes of student evaluation, faculty members shall not be entitled to the identity of the student respondent unless such evaluation is utilized as a basis for termination or other disciplinary action."

This instrument was administered for all full-time and part-time faculty employed in the day and continuing education divisions at one community college in the system. Only full-time day division faculty throughout this particular 15 community college system were bound by contractual obligation

to undergo the evaluation. The descriptive statistical analysis presented in this section was tabulated from data collected from division of continuing education students who evaluated their teachers. Scores for each item were reported on the basis of percentage of students responding to each scale choice; for example, if 60% of the students thought that the instructor "almost always" held classes as scheduled and 20% indicated that the instructor "frequently" held classes as scheduled, and so forth for the remaining choices, then the instructor's evaluation results might look something like this:

ALWAYS	FREQUENTLY	SOMETIMES	INFREQUENTLY	HARDLY EVER	NO RESPONSE
60.00%	20.00%	10.00%	10.00%	0.00%	0.00%

Table 1 provides elementary statistics for summed values of "always" and "frequently" responses. In other words, percentage scores under "almost always" and "frequently" for each item were added together and this sum was combined with that of the 139 other scores to calculate the mean score for that item. The scores of

instructors evaluated by fewer than ten students were not entered into the calculations. The information contained in Table 1 and graphically represented in the curves of frequency distributions in Figure 1 portray the instrument's inability to discriminate superior, average, and poor college teachers. Thus, statistics in Table 1 show that with the exceptions of items 5, 6, and 9, on average more than 85% of the students evaluating instructors chose to rate them on the highest two performance scales. One would like to know whether the instructors being rated are as superior as the results indicate or were they the fortunate beneficiaries of a seriously flawed instrument. Clearly, the more realistic approach is to investigate the shortcomings of the instrument and attempt to determine why the instrument failed for the most part to distinguish superior from average and poor teachers. There is no claim implied that the analysis of this instrument's efficacy appearing here is complete.

TABLE I

MEAN AND STANDARD DEVIATION PERCENT OF
RESPONDENT SCORES ON FACULTY EVALUATION
N = 140

ITEM	1	2	3	4	5	6	7	8	9	10	11	12
MEAN	87.657	87.701	92.055	86.788	82.435	78.532	96.472	92.799	78.088	88.103	86.677	90.604
S.D.	16.695	14.283	10.732	14.167	17.602	17.523	9.106	16.981	18.010	13.485	13.201	11.244

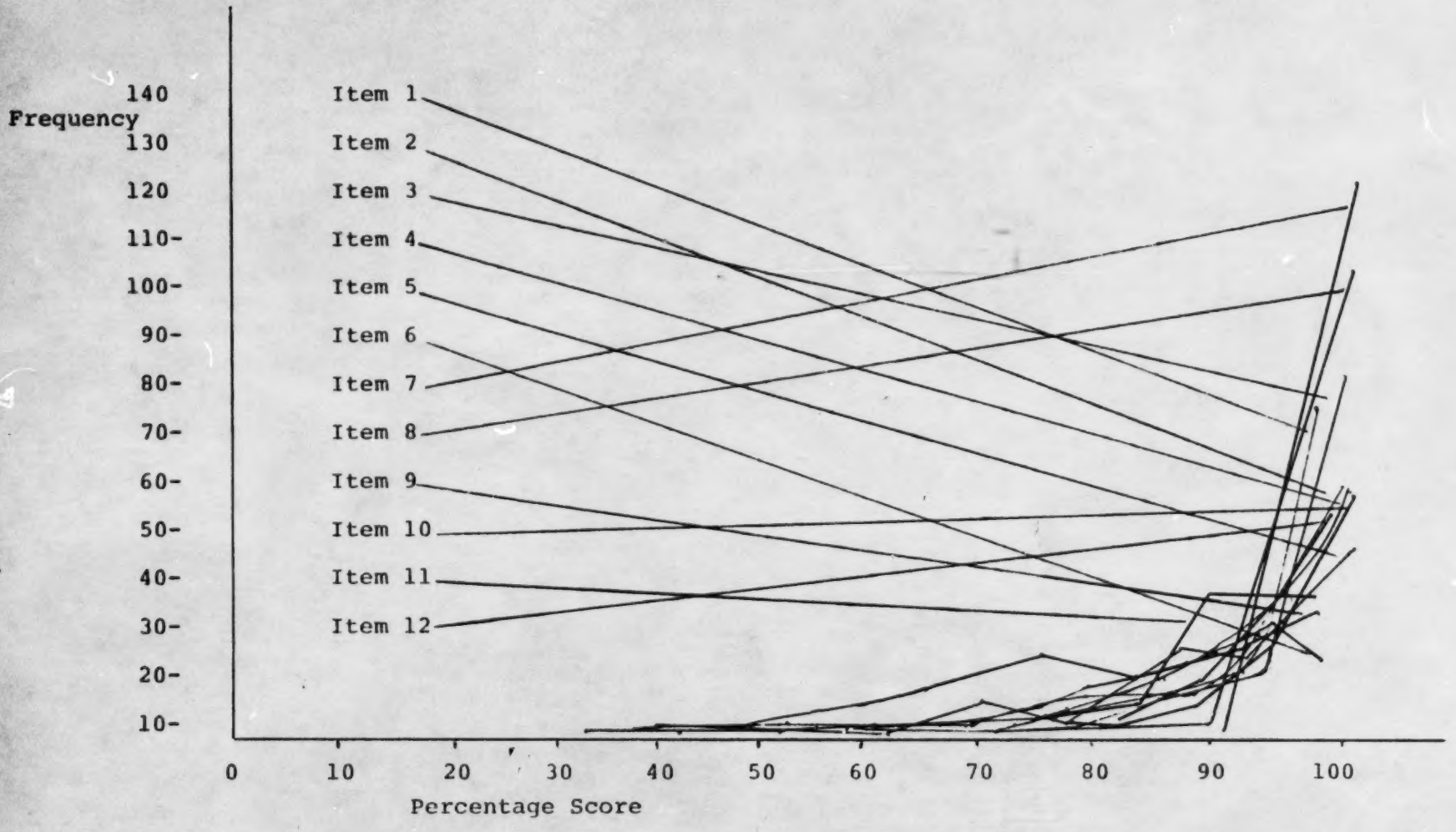

Figure 1. Frequency Distribution Curves of Instructors' Evaluation Scores for 12 Item Rating Form

Table 1 and Figure 1 are merely indicative of serious deficiencies in discriminative power and that the succeeding comments may provide reasons for this impotency. The most fundamental reason, however, for the instrument's technical and practical problems rests inherently with the process which bore it - the collective bargaining process. A more scientific approach would have been incongruous with the methods used in collective bargaining but may have avoided the pitfalls explored in the remainder of this section.

In the service of clarity and brevity and perhaps at the expense of a complete discussion the major weaknesses of this student evaluation of faculty instrument and its administration will be listed:

1. Raters were not prepared prior to making observations, i.e. they were not notified of performance criteria.
2. Each choice in the response scale ranging from "almost always" to "almost never" is an indefinite designation of frequency; hence ambiguous. Quantitively, one wonders what the difference is, in the minds of students, between something happening almost

always and frequently; likewise infrequently and almost never.

3. The uniform application of one response scale to all items contributes to a response set.
4. Only twelve items were used to measure the complex behavior and interaction of college teaching. Additionally, on the face of it there appears to be no more than two and in some cases only one item called upon to measure discrete behavior.
5. Of the twelve items, four (4,7,8,10) could be considered to be concerned with the mechanics of teaching rather than effective teaching itself.
6. The lack of total anonymity for students coupled with the administration of this evaluation instrument before final grades were issued may have augmented the tendency to be lenient by unwitting intimidation.
7. Items 5, 8 and 9 are highly ambiguous.
8. Item 6 attempts to measure a covert trait.
9. Since no tests of validity or reliability were employed, there is no assurance that

the instrument measures what the negotiators wanted it to measure or that what was judged was not a fleeting piece of behavior rather than an enduring attribute.

10. Since there was no scientific analysis of the instrument, there appear to be few safeguards developed against such testing errors as: perfunctoriness, response-set, "halo"-effect, leniency, or acquiescence.

RECOMMENDATIONS

The evidence offered in the foregoing discussion supports the conclusion that a student rating instrument designed to measure the classroom performance of college teachers ought not to be developed at the collective bargaining table. Simply stated: reality is not readily measured by consensus. The process of collective bargaining is not likely to become more scientifically oriented. Yet, it is equally unlikely that organized faculty and management will cease to include student ratings as an integral part of the evaluation process. Centra (2) indicates that dropping the results of student ratings for certain faculty would be perceived as the loss of an

objective measurement which works to counteract administrative favoritism. Abandoning the results of these ratings for deans and department chairmen, who face fewer opportunities to grant promotions and tenure, would result in less evidence available on which to base their recommendation.

Student ratings will be with us in the foreseeable future and will, for unionized faculty, be subject to negotiations as a condition of employment. What intelligent course of action remains in light of the apparent irreconcilable differences between the goals of collective bargaining and scientific measurement? There are at least two options available: (1) the parties can agree to use a reliable and validated instrument developed by a testing service or other institution, or (2) methodically develop an instrument away from the bargaining table. In the latter option the parties to the negotiations could initially enumerate the different categories of teaching behavior they wanted measured and then give the task of developing a student rating instrument to a select team of individuals skilled in the science of psychological testing. The team, after completing its work, could submit their

product to the negotiators for approval. In this manner, perhaps, both faculty and administrators would be more confident in the process of evaluation and avoid wasting a great deal of time, money, and energy on unworthy instruments.

REFERENCES

1. Carr, Robert K. and Van Eyck, Daniel K. Collective Bargaining Comes to the Campus. Washington, D.C.: American Council on Education, 1973

2. Centra, John A. "The Student as Godfather? The Impact of Student Ratings on Academia" Educational Researcher 2 (1973).

3. Cronbach, Lee J., Essentials of Psychological Testing (3rd ed.) New York: Harper & Row, 1970.

4. Downie, N.M., Fundamentals of Measurement (2nd ed.) New York: Oxford University Press, 1967.

5. Kemerer, Frank R. and Baldridge, J. Victor, Unions on Campus. San Francisco: Jossey-Bass, 1975.

6. Miller, Donald W. "Dangers of Using Student Evaluation for Administrative Purposes" Collegiate News and Views. 31 (1978): 2-3.

7. Seldin, Peter. How Colleges Evaluate Professors. Croton-on-Hudson, New York: Blythe-Pennington, LTD., 1975.

8. Simkin, William E. "Fact-Finding - Its Values and Limitations" 61 Government Employment Relations Reporter. (1970)

9. Thorndike, Robert L. and Hagen, Elizabeth, Measurement and Evaluation in Psychology and Education (2nd ed.) New York: John Wiley & Sons, 1965.

10. Tice, Terrence N. (ed.) Faculty Bargaining in the Seventies Ann Arbor: The Institute of Continuing Legal Education, 1973.

FEB 26 1982
ERIC Clearinghouse for Junior Colleges
96 Powell Library Building
University of California
Los Angeles, California 90024